Web usage mining for the new seismic hazard map of Italy

Giuliana Rubbia
Istituto Nazionale di Geofisica e Vulcanologia, Sezione di Milano
via Bassini 15, 20133 Milano, Italy
e-mail rubbia@mi.ingv.it

keywords: web usage mining, web server statistics, hazard map

Extended abstract
Following the Ordinanza PCM 20 marzo 2003, n. 3274, a new map of seismic hazard of Italy has been produced by a national Working Group at INGV, Italian National Institute of Geophysics and Volcanology.
To fulfil the criterion of transparency of the compilation process, in July 2003 a dedicated web site was set up to disseminate information about the Ordinance itself, and make publicly available work in progress, procedures, input data and final results (http://zonesismiche.mi.ingv.it).
An analysis of the web usage is presented for the period of one year and in particular after the publication of final results.

Web usage mining identifies a complex procedure by which statistical methods and data mining technologies are employed in order to extract implicit, previously unknown and potentially useful information from *web data*. Web data are classified by Srivastava et al. (2000) into four types: a) content, b) structure, c) usage and d) user profile, where a) refers to the real data in the web site and b) describes its organization; usage data, registered in the log file, include IP address, date and time of access, requested file, and other parameters of the connection between the client of the user and the web server; user profile data provide demographics information, such as name, country, profession and interests of users; they can be acquired through questionnaires or registration forms or be extracted by analyzing the log file.

Eirinaki & Vazirgiannis (2003) present a survey of the most important commercially available products and research efforts in the field of using data mining technologies for web personalization, that is for the process of customizing a web site to the needs of specific users taking advantage of the knowledge acquired from the analysis of usage data in correlation with content, structure and user profile data.
Statistical analysis of data stored in the log file is the most common and simple method; it can be performed by using one of several existing log analysis tools.
More advanced methods are used to cluster and classify data in order to extract navigation patterns, groups of pages and groups of users, and relationships between them (Srivastava at al., 2000; Mobasher, 2004). Spilioupoulu & Pohle (2001) define the success of a site through the analysis of navigation patterns; in the field of e-commerce, new metrics are defined to measure the conversion of web site visitors into customers (Cutler & Sterne, 2000; Teltzrow & Berendt, 2003).

For the web site dedicated to the new hazard map, web usage mining was mainly performed on a statistical basis. Two very popular software packages for web server statistics have been used: Analog (http://www.analog.cx/), used for example by USGS for monthly report, and Wusage (www.boutell.com/wusage).
As first step, typical statistical indicators were derived, such as number of requests and number of visitors within a given period, domain distribution requests, top referring sites, search words, top pages visited, and so on.
In parallel, special attention was devoted to understand which categories of users visit the web site. User profile data were inferred by analyzing the log file. The owner of an

IP address was determined from publicly accessible domain registration databases, such as Italian Registration Authority and RIPE; further information about the owner were derived from other sources, for example browsing the web site of the owner of interest, as suggested in (Pitkow, 1997). Single IP addresses (e.g. xyz.unimi.it) have been grouped into “entities” (University of Milano) and consequently into categories (“Universities and Research Bodies”). The resulting database was progressively populated; in one year of activity 20,000 IP addresses were collected and grouped into 2,000 entities; it was possible to assign 1,500 of them to 30 categories. Moreover, more reliable demographics were collected by asking the actual visitors of a web page, for which a registration form was inserted.
Finally, relationships between groups of visitors and group of pages were investigated. Requests distribution of categories of users are provided for those information published at the end of the project, including earthquake catalogue, seismic zonation, hazard map and its values.

According to this process, the following considerations were drawn.

The web site of the new seismic hazard map of Italy has been visited on a regularly basis. Both visitors and requests increased after the publication of final results of the project. Peaks were registered in correspondence with intermediate milestones of the project, advertisements of new information available, and on the occasion of seismic events.
A rough estimation of demographics, based on categorization of IP addresses, revealed that visitors matched the expected audience: they were both contributors and end-users. Several categories of users were found: universities and research bodies, regional and local administrations, ministries, environmental agencies and consortia, health services, military, technicians’ offices, financial groups, constructions companies, insurance companies, power suppliers, schools, newspapers and media.
The comparison of requests for earthquake catalogue, seismic zonation, hazard map and its values, aimed to see if some categories are more interested in one piece of information, was affected by a high percentage of requests from Internet Service Providers. Nevertheless, a more reliable analysis was possible for visitors that registered to the web site and downloaded PGA values. These visitors include administrators, professionals, researchers, and companies which activities are related to infrastructures and environment. Although on a limited set of data, the percentage of requests from universities and research bodies (31%) can be compared with the sum of requests made by professionals, geologists and engineers (18%) and by technicians’ offices (14%); local administrations follow (20%).

The presented analysis gave some indications about the response to the information dissemination activity. Nevertheless, some developments are envisaged: a) to improve procedures of collecting user profile data, through registration forms or questionnaires (see for example EMSC, 2004); b) to explore and consequently adopt software packages and/or research prototypes which go beyond statistical analysis, in order to better characterize navigational behaviour of users, understand and better serve their needs.